**מגישים :**

**עומר פלד – 315110015**

**חן סקנדרני – 206508541**

**שם הפרוייקט :** Danger of fires

**רקע על NIFC:**

מרכז האש הבין-ארגוני הלאומי הוא מיזם של מספר סוכניות פדרליות בארצות הברית, מטרת ה-NIFC היא לספק מענה לאומי מתואם לשריפות ואירועי חירום אחרים.

ה-NIFC משמש כמרכז התמיכה הלאומי לכיבוי אש באדמות פרא. הוא מספק תמיכה לוגיסטית ותפעולית, כולל כוח אדם, מטוסים, ציוד ומודיעין לסוכנויות מקומיות, מדינתיות ופדרליות המגיבות לשריפות ותקריות אחרות. המרכז גם שומר ומפיץ מידע לאומי על שריפות פרא, כגון תחזיות מזג אוויר שריפות, דירוגי סכנת שריפה ותחזיות התנהגות שריפות.

**שאלת המחקר :** האם ניתן לחזות מהו הגורם המרכזי לפריצת שריפה?

**למה בחרנו בנושא :**

בחירת הנושא הייתה לא פשוטה, חיפשנו נושא שמעניין את שנינו לאחר התלבטות ומחקר על הרבה נושאים לחזות כגון: (נדלן, אולימפיאדה, גובה הכנרת, כמות תאונות בישראל).

לצערנו אחרי מחקר ממושך לא מצאנו מספיק מידע על הנושאים מכיוון שלא הייתה יכולת ממשית לחזות עמודת מטרה מסויימת.

בסופו של דבר עלה לנו רעיון שאפשר לנסות לחזות גורם מסויים לאסונות טבע, שריפות בפרט.

החלטנו לבחור בנושא השריפות שעניין את שנינו ומצאנו עליו מספיק מידע על מנת לבצע ניקיון כמו שצריך ולמידת מכונה טובה.

**תהליך עבודה ושלבי עבודה :**

**הרכשת מידע שריפות :**

מצאנו אתר מצויין לשאיבת נתוני שריפות בארצות הברית במהלך השנים 2014-2023, כל המידע היה בתוך טבלה עם 250 אלף שורות כאשר יש גישה ל25 שורות בכל פעם.

חיפשנו דרך להגיע לכל האלמנטים המצויים בטבלה בלי לרדת עד סוף הטבלה(250 אלף שורות...)

בעזרת סלניום(SELENIUM), השתמשנו בfirefox driver על מנת לעשות גלילה למטה עד שכל הטבלה וכל האלמנטים יהיו זמינים להרכשה בעזרת Beautiful Soup.

מכיוון שהאתר היה איטי, לקח שבוע של הרצת הקוד על מנת שהסלניום יוכל לרדת לסוף הטבלה.

ברגע שהסלניום סיים את עבודתו, ביצענו 2 פונקציות אשר רוכשות את המידע מהטבלה לפי שורה ועמודה והופכות את המידע שרכשנו לקובץ .CSV

**מידע מזג אוויר :**

לאחר סיום ההרכשה של כל השריפות, רצינו להשתמש בWeather api על מנת לרכוש נתוני מזג האוויר על פי מיקום וזמן השריפה.

נקלענו בקושי מרכזי: כל אתר עם API זמין לשאיבת נתוני מזג האוויר בארצות הברית דרש כמות כסף בתמורה לנתונים, לאחר מכן הבנו שאנחנו חייבים לצמצם את כמות נתוני מזג האוויר בצורה משמעותית על מנת שנוכל לרכוש מספיק נתונים ובו זמנית לא לפשוט רגל.

האתר הכי זול שמצאנו נתן לנו לשאוב 1000 שורות של נתוני מזג האוויר ביום באופן חינמי.

החלטנו לשנות גישה ולקחת מיקום מזג אוויר כללי בארצות הברית לפי טווח התאריכים של כל השריפות.

מ2014 ועד היום, כלומר בערך 3000 ימים, אותם הוצאנו 1000 בכל יום, סהכ 3 ימים וסידרנו את כל הנתונים בcsv יחיד.

בסופו של דבר הגענו ל2 קבצי CSV , 1 של שריפות , 1 של מזג האוויר- נצטרך לאחד ביניהם.

**שלב מיזוג המידע :**

קושי מרכזי: פורמט התאריכים בשני הקבצים היה שונה, ורצינו לאחד את שני הקבצים בעזרת התאריכים, בעקבות זה בנינו פונקציה אשר הופכת את כל התאריכים בשני הקבצים לפורמט דומה על מנת שנוכל לאחד ביניהם לאחר ניקוי כל קובץ בנפרד.

**ניקוי מידע השריפות:**

ביצענו מס פונקציות שבעזרתן ניקינו את המידע:

1. טיפלנו בשורות משוכפלות.
2. טיפלנו בערכים חסרים.
3. מתוך 96 עמודות סיננו ל8 עמודות רלוונטיות בלבד.
4. שינינו את סוג העמודה של גורם השריפה ממחרוזות לקטגוריות 1 - בן אדם, 2- טבע , 3 - לא ידוע , 4 - לא סווג.
5. שינינו את הפורמט של התאריכים.
6. טיפלנו בערכים חריגים(OUTLIERS) בעמודות מספריות רלוונטיות(נקודות ציון)

הוספנו עמודה מטרה בינארית בשם CausedByWeather כאשר 1 זה אומר שהשריפה נגרמה על ידי מזג האוויר ו0 שריפה נגרמה על ידי בן אדם או על ידי גורם לא ידוע.

**ניקוי המידע המאוחד ( מזג האוויר ושריפות ביחד ) :**

ביצענו בדיקה ידנית על הקובץ המאוחד בעמודות המזג אוויר: (טמפרטורה, לחות, אינדקס קרינה, מהירות רוח), בדקנו את טווחי הנתונים האפשריים בארצות הברית בפרט ובכך גילינו שאין נתונים חריגים וכל הנתונים שלקחנו מתאימים לשריפות המסויימות.

ווידאנו שאין ערכי NAN.

בסיום השלב הזה, הגענו לקובץ מאוחד ונקי עם כל המידע הרצוי, לשלב הבא

גודל הקובץ הסופי לאחר סינונים:

כמות שורות: 108,068 **ייצוג :** שריפה

כמות עמודות: 14

**ייצוג:**

1. תאריך גילוי השריפה
2. מזהה שריפה
3. שם השריפה
4. קו אורך
5. קו רוחב
6. תאריך סיום השריפה
7. גורם השריפה
8. היקף השריפה(כמה שטח נשרף)
9. טמפרטורה
10. לחות
11. מהירות רוח
12. כיוון הרוח
13. אינדקס קרינה
14. עמודה בינארית- נגרם על ידי מזג האוויר

**ויזואליזציה :**

על מנת לראות מידע על הנתונים שישפיעו בסופו של דבר על הלמידת מכונה, נעזרנו בספריות: matplotlib וseaborn כדי להציג את הנתונים בגרפים.

**גרפים חד מימדיים :**

דבר ראשון רצינו למצוא את הגורמים המרכזיים לגרימת שריפות על ידי מזג האוויר, הצגנו את כל העמודות שלנו בגרף BAR כדי לראות את התדירות של כל אחד מהעמודות הבאות: טמפרטורה, לחות ומהירות רוח. בנוסף ראינו את התפלגות הנתונים:

הטמפרטורה השכיחה ביותר לשריפה הוא 25 מעלות.

טווחי מזג האוויר האפשריים : [-9.8 , 32.1]

הלחות השכיחה ביותר לשריפה היא 65 אחוזים.

טווחי מזג האוויר האפשריים : [19.8 , 97.9]

מהירות הרוח השכיחה ביותר לשריפה היא 20 קמ"ש.

טווחי מזג האוויר האפשריים : [4.3, 69]

לסיום, הצגנו גרף פאי של התפלגות גורמי השריפה באחוזים, מתחלק ל: בני אדם, טבע, לא ידוע.

**גרפים דו מימדיים :**

המטרה העיקרית היא לראות את הקשר בין גורמים משפיעים של מזג האוויר לגורם השריפה.

1. השתמשנו בscatterplot על מנת להראות את ההשפעה של מהירות הרוח והטמפרטורה גורמות ביחד על גורם השריפה- קל לראות שכאשר מהירות הרוח נמוכה יחסית והטמפרטורה גבוהה, ההסתברות יותר גבוהה שהשריפה נגרמה על ידי טבע/מזג האוויר.
2. השתמשנו בscatterplot על מנת להראות את ההשפעה של הלחות והטמפרטורה גורמות ביחד על גורם השריפה- קל לראות שכאשר הלחות גבוהה יחסית והטמפרטורה גבוהה, ההסתברות יותר גבוהה שהשריפה נגרמה על ידי טבע/מזג האוויר.
3. השתמשנו בbar graph להראות את הקשר בין אינדקס הקרינה לתדירות כמות השריפות אשר נגרמו על ידי גורמים שונים וכך גילינו שככל שאינדקס הקרינה עולה, ככה תדירות השריפות אשר נגרמו על ידי הטבע/מזג האוויר עולות בהתאמה, והשריפות שנגרמו על ידי בני אדם יורדות בהתאמה.

**גרפים תלת מימדיים :**

בגרף הבא רצינו להציג את הקשר בין שלושת הגורמים החזקים ביותר שמשפיעים על השריפות ולכן בחרנו להשתמש בscatterplot 3d שבוחן את הקשר בין טמפרטורה, אינדקס קרינה ולחות.

**למידת מכונה :**

עמודת המטרה שלנו היא עמודה בינארית ולכן אנו צריכים להשתמש בלמידה מונחית בקלסיפיקציה.

הקלסיפיקציה היא תהליך של חיזוי משתנה מטרה עם ערך קטגורי. ישנם מספר מודלים קלסיפיקציה שונים, והבחירה של מודל מתבססת על תכונות הנתונים והקשר ביניהם.

יש כמה אלגוריתמים אשר יכולים להביא את התוצאה המתבקשת, אנחנו בחרנו ברגרסיה לוגיסטית.

רגרסיה לוגיסטית הינו אלגוריתם אשר מודל את הקשר בין עמודות המידע לבין עמודת המטרה הבינארית ומאפשר להעריך את ההסתברות של התוצאה.

**סיכום:**

לפני המחקר הסקנו ששריפות הנגרמות על ידי הטבע נגרמות רק בגלל טמפרטורות גבוהות והתחממות כדור הארץ אך לאחר מחקר מעמיק בנושא גילינו שלא רק הטמפרטורה היא גורם לשריפה.

הגענו למסקנה שהגורם המרכזי לשריפות הנוצרות על ידי הטבע הוא ההשפעה של הלחות הגבוהה, מזג האוויר הגבוה, מהירות רוח יחסית נמוכה, ואינדקס קרינה גבוה.

לסיום, נתקלנו בפרוייקט בקשיים רבים בכל התחומים, אך בעזרת שימוש באינטרנט וביוטיוב למדנו את הדברים הנוספים (selenium) והצלחנו לבצע את כל השלבים הרצויים.

הקורס העשיר את הידע שלנו בפייתון ובמדעי הנתונים בפרט.

למדנו את חשיבות הלמידה העצמית וראינו שלא משנה בכמה בעיות ניתקל תמיד אפשר לפתור אותם בעזרת מתרגלים במכללה ומרצים.